The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2385–2402
DOI: 10.1214/11-AOAS492

MULTIVARIATE BAYESIAN SEMIPARAMETRIC MODELS FOR AUTHENTICATION OF FOOD AND BEVERAGES[1]

By Luis Gutiérrez and Fernando A. Quintana[2]

Universidad de Chile and Pontificia Universidad Católica de Chile

Food and beverage authentication is the process by which foods or beverages are verified as complying with its label description, for example, verifying if the denomination of origin of an olive oil bottle is correct or if the variety of a certain bottle of wine matches its label description. The common way to deal with an authentication process is to measure a number of attributes on samples of food and then use these as input for a classification problem. Our motivation stems from data consisting of measurements of nine chemical compounds denominated Anthocyanins, obtained from samples of Chilean red wines of grape varieties Cabernet Sauvignon, Merlot and Carménère. We consider a model-based approach to authentication through a semiparametric multivariate hierarchical linear mixed model for the mean responses, and covariance matrices that are specific to the classification categories. Specifically, we propose a model of the ANOVA-DDP type, which takes advantage of the fact that the available covariates are discrete in nature. The results suggest that the model performs well compared to other parametric alternatives. This is also corroborated by application to simulated data.

1. Introduction. Food and beverage authentication is the process in which foods or beverages are verified as complying with its label description [Winterhalter (2007)]. From the viewpoint of consumers' acquisition, the mislabeling of foods represents commercial fraud [Mafra et al. (2008)]. On the other hand, producers and sellers could have problems if their products are mislabeled. Food authentication is important for foods and beverages of high commercial value, like honey, wines or olive oil, because their prices depend of their quality, variety or origin. It is then important to uncover unscrupulous sellers who decide to increase their profit by adulterating

Received November 2010; revised June 2011.
[1]The anthocyanin profiles were obtained in the frame of the FONDEF Grant D00I1138.
[2]Supported in part by Grant FONDECYT 1100010.
Key words and phrases. Classification, dependent Dirichlet process, wines.

these products with similar but lower quality substances. Misleading labeling might also have negative health implications, especially when the food has undeclared allergenic compounds.

Because of the growing demand from consumers of clarity and certainty in food origins and contents, the importance of food authentication has substantially increased in recent years. Many analytical tools and methods used for authenticity have been consequently developed. In particular, there is a very active area of research on the determination of chemical markers for classification and/or authentication of wines. Anthocyanin profiles are known to be specially useful for the purpose of wine variety authentication. See, for example, Eder, Wendelin and Barna (1994), Berente et al. (2000), Holbach, Marx and Ackerman (2001), Revilla et al. (2001), Otteneder, Marx and Zimmer (2004) and von Baer et al. (2007).

Data analysis methods for authentication purposes have been developed mainly outside the statistics fields, and most of them are exploratory techniques designed to deal with multivariate data sets. Probabilistic modeling for discrimination and authentication purposes was proposed by Brown, Fearn and Haque (1999), who used Bayesian methods to discriminate 39 microbiological taxa using their reflectance spectra. More recently, Dean, Murphy and Downey (2006) used a Gaussian mixture model with labeled and unlabeled samples, with application to the authentication of meat samples from five species, and the geographic origin of olive oils. Toher, Downey and Brendan (2007) compared model-based classification methods such as Gaussian mixtures with partial least squares discriminant analysis, considering samples of pure and adulterated honey.

We propose a model-based procedure to solve the authentication problem of foods and beverages. The motivation comes from a data set consisting of measurements of nine chemical compounds denominated Anthocyanins, obtained from samples of Chilean red wines of grape varieties Cabernet Sauvignon, Merlot and Carménère. We propose a semiparametric Bayesian model that allows us to define a flexible distribution G for the joint measurements. The model has the advantage of not having to assume any parametric form, which may be particularly difficult to check in multivariate cases. Increased flexibility is added by allowing G to be formulated under the formalism of dependent random probability measures as in De Iorio et al. (2004). An interesting aspect of the proposed approach is the need to extend previous univariate semiparametric models as in De la Cruz-Mesía, Quintana and Müller (2007) to the multivariate case.

The rest of the paper is organized as follows. We first present the wine data set and the related authentication problem in Section 2. In Section 3 we give a brief theoretical background about Bayesian semiparametric models and dependent Dirichlet processes, and discuss our approach to the authentication problem. In Section 4 we present the model, which is an extension

of the univariate semiparametric Bayesian linear mixed model [Dey, Müller and Sinha (1998)] to the multivariate case. In Section 5 we illustrate the performance of the proposed model in a simulated data set. In Section 6 we apply the model to authenticate red wines samples based on their anthocyanin profile. The paper concludes in Section 7 with a discussion and final remarks.

2. The motivating data set. We consider a data set consisting of measurements of concentrations of nine anthocyanins on samples of Chilean red wines. Anthocyanins are a group of chemical compounds present in red wine, which confer to this beverage its characteristic red color and are transferred from the grape skins to wine during the winemaking process. The data set includes the grape variety for each sample *as declared by the producer*, the year of harvest and the geographical origin or valley. The grape varieties in the data set are Cabernet Sauvignon (228 samples), Carménère (95 samples) and Merlot (76 samples). These samples form a data set with mixed wine types, which represents the most abundant red grape varieties cultivated in Chile across different valleys. The data are unbalanced, and there are combinations of variety and location for which no observations are available. All wine samples came directly from wineries located in the valleys of Aconcagua, Maipo, Rapel, Curicó, Maule, Itata and Bío-Bío in Chile. They correspond to the vintages 2001, 2002, 2003 and 2004. Anthocyanin determination was made by reverse phase High Performance Liquid Chromatography (HPLC), a Chromatographic technique that can separate a mixture of compounds and is used in analytical chemistry to identify, quantify and purify the individual components of complex mixtures, like wines and others beverages or foods. The analytical chemistry procedure was based on the method described by Holbach, Marx and Ackermann (1997), Otteneder et al. (2002) and by the International Organization of Vine and Wine (OIV) [OIV (2003)], with some minor modifications. More details about anthocyanin determination for this data set can be found in von Baer et al. (2005, 2007). A main concern for these data is the authentication of grape variety using the log-concentrations of the following anthocyanins: delphinidin-3-glucoside (DP), cyanidin-3-glucoside (CY), petunidin-3-glucoside (PT), peonidin-3-glucoside (PE), malvidin-3-glucoside (MV), peonidin-3-acetylglucoside (PEAC), malvidin-3-acetylglucoside (MVAC), peonidin-3-coumaroylglucoside (PECU) and malvidin-3-coumaroylglucoside (MVCU). To do so, we propose in Section 4 a multivariate linear mixed model that attempts to characterize the variability in anthocyanin log-concentrations in terms of variety and valley of origin. We also point out that we ignore vintage year in our development. The pragmatical reason for this is that by doing so we may easily incorporate data from new years as they become available, without the need to modify the model.

In support of this choice, we refer to Gutiérrez et al. (2011) who used the year of harvest as a continuous predictor when proposing a Bayesian parametric model for the same data. The idea was to overcome this very same limitation. Yet, the effect of vintage year was negligible in that context.

3. Some background material. Semiparametric models have both parametric and nonparametric parts, the distinction between these being that the parameters belong to a finite and infinite-dimensional space, respectively. Semi- and nonparametric Bayesian models are used mainly to avoid critical dependence on parametric assumptions. An important application of such modeling strategy is to random effects distributions in hierarchical models, where often little is known about the specific form of such distributions [Müller and Quintana (2004)]. To handle the nonparametric part of the model, we need to define a random measure on the space of distribution functions. The most popular such choice is the Dirichlet process (DP) [Ferguson (1973)].

In a food authentication context scenario, we need to build a model that adequately accounts for all the problem-specific features. In the specific case of our motivating data set, it is reasonable to think of wines coming from the same valley as being correlated, because soil and weather conditions are similar within a given valley. The usual (and simplest) way to induce a correlation structure is by incorporating random effects or sample specific parameters in a model. Let α_i denote the random effects and let z_i be a categorical covariate with k levels (e.g., k different regions of origin). We could assume a single nonparametric prior on α_i for all samples, without reference to the levels of z_i. Alternatively, we could consider differences by putting k independent priors on α_i. These two extreme modeling strategies imply that $G_{z_1} = \cdots = G_{z_k}$ for the former and $G_{z_1}, \ldots, G_{z_k}$ to be mutually independent for the latter. MacEachern (1999) proposes a modeling strategy, the Dependent Dirichlet Processes (DDP), that allows the set of random effects distributions to be similar but not identical to each other. MacEachern (1999) defines a nonparametric probability model for G_z in such a way that marginally, for each $z = z_j$ $(j = 1, \ldots, k)$, the random measure G_z follows a DP. In this context, the DP representation proposed by Sethuraman (1994) is quite useful. Sethuraman's representation establishes that any $G \sim \mathrm{DP}(M, G_0)$ can be represented as an infinite mixture of point masses:

$$
\begin{aligned}
G(\cdot) &= \sum_{h=1}^{\infty} w_h \delta_{\mu_h}(\cdot), \qquad \mu_h \overset{\text{i.i.d.}}{\sim} G_0, \\
w_h &= U_h \prod_{j<h} (1 - U_j) \qquad \text{with } U_h \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, M).
\end{aligned} \tag{1}
$$

The key idea behind the DDP is to introduce dependence across the G_z measures by assuming the distributions of the point masses to be depen-

dent across different levels of z (i.e., μ_{zh}), but still independent across h. If the weights are assumed to be the same across z, the dependent probability measure can be represented as $G_z(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_{zh}}$. The last idea was used by De Iorio et al. (2004) in the construction of an ANOVA-DDP type model. A similar approach was used in spatial modeling by Gelfand, Kottas and MacEachern (2005), who proposed a Gaussian process for the atoms, Caron et al. (2006) in times series, De la Cruz-Mesía, Quintana and Müller (2007) in classification, De Iorio et al. (2009) in survival analysis and, recently, by Jara et al. (2010) who proposed a Poisson–Dirichlet process for the analysis of a data set coming from a dental longitudinal study. Griffin and Steel (2006) point out that letting only the atoms depend on covariate values may lead to certain problems when points in the domain are far from the observed data. They propose an approach that avoids this by locally updating the process and inducing dependence in the weights through distance-based similarities in the ordering of atoms, through viewing the atoms as marks in a point process. Other works where covariate dependence is introduced in the weights are Dunson, Pillai and Park (2007) and Dunson and Park (2008). Müller, Erkanli and West (1996) considered a completely different approach for inducing dependence in G. They used a DP mixture of normals for the joint distribution of y and z, and then focused on the implied conditional density of y given z for estimating the mean regression function. A recent reference about nonparametric Bayesian statistics, DDP models and their applications can be found in Hjort et al. (2010).

The almost sure discreteness of the Dirichlet process makes it inappropriate as a model for a continuous quantity y. A standard procedure for overcoming this difficulty is to introduce an additional convolution so that

$$H(y) = \int f(y \mid \theta)\, dG(\theta) \qquad \text{with } G \sim \mathrm{DP}(M, G_0). \tag{2}$$

Such models are known as DP mixtures (DPM) [Antoniak (1974)]. The mixture model (2) can be equivalently written as a hierarchical model by introducing latent variables θ_i and breaking the mixture as

$$y_i \mid \theta_i \sim f(y_i \mid \theta_i), \qquad \theta_i \sim G \quad \text{and} \quad G \sim \mathrm{DP}(M, G_0). \tag{3}$$

For the majority of food authentication problems the responses are continuous multivariate and covariates are discrete. This is the case for the data described in Section 2. Thus, we will adopt the popular semiparametric modeling strategy that consists of introducing dependence in the random effects distribution and then adding a convolution with a continuous kernel. The ANOVA-DDP approach of De Iorio et al. (2004) is a natural way to build the desired dependence into the model, as will be discussed below in Section 4. We remark here that a model that defines dependence in terms of distances would not be appropriate for an authentication problem with categorical covariates, as is our case.

4. The model. We first note that due to the multivariate nature of many authentication problems (which is also the case of the wine data), it would not be appropriate to treat the individual responses in an univariate way.

We assume that the ith response vector is related to the covariates in a linear way. Furthermore, we assume that there are fixed and random effects in the model. The model for the ith unit in the uth group or class in our classification context is thus assumed to be given by

$$\begin{aligned}(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \qquad i = 1, \ldots, n_u, u = 1, \ldots, m,\\ \theta_{iu} &\sim H_z(\theta_{iu}),\\ H_z(\theta) &= \int N(\theta \mid z\alpha, \tau)\, dG(\alpha),\\ G &\sim \mathrm{DP}(M, G_0),\end{aligned} \tag{4}$$

where y_{iu} is a vector of responses in R^p, B is a $p \times q$ matrix of fixed effects, x_{iu} is a vector of covariates in R^q, θ_{iu} is a $p \times 1$ vector of unit-specific random effects, z_{iu} is a $p \times pk$ design matrix for random effects and α is a $pk \times 1$ vector of latent variables that define the random effects. For classification purposes, the subscript u denotes the group or class. Model (4) implies that $H_z(\theta) = \sum_{h=1}^{\infty} w_h N(\theta \mid z\alpha_h, \tau)$ is an infinite mixture of normal distributions. As usual in mixture models, posterior simulation proceeds by breaking the mixture in (4) by introducing latent variables α_i:

$$\theta_{iu} = z_{iu}\alpha_i + \eta_i, \qquad \alpha_i \sim G, \qquad G \sim \mathrm{DP}(M, G_0) \quad \text{and} \quad \eta_i \sim N_p(0, \tau). \tag{5}$$

By simplicity, we choose a multivariate normal model for the base measure $G_0 \equiv N_{pk}(0, R)$ and as usual in this context, we assume prior independence for all remaining parameters. The prior distribution for matrix $B = [\beta_1, \beta_2, \ldots, \beta_q]$ is assumed to be independent by columns, that is, $\beta_1, \beta_2, \ldots, \beta_q$ are mutually independent with distribution given by

$$\beta_1, \ldots, \beta_q \sim N_p(\beta_{0j}, \Lambda), \qquad j = 1, \ldots, q. \tag{6}$$

The prior distributions for the variance–covariance matrices Σ_u, $u = 1, \ldots, m$, and τ are given by

$$\Sigma_1, \ldots, \Sigma_m \sim IW_p(\nu_0, Q_0), \qquad \tau \sim IW_p(\gamma_0, \Phi_0). \tag{7}$$

We complete the Bayesian formulation of model (4) by specifying the prior for hyperparameters R, $\beta_{01}, \ldots, \beta_{0q}$, Λ and M as

$$R \sim IW_{pk}(r_0, R_0), \qquad \beta_{01}, \ldots, \beta_{0q} \sim N_p(\alpha_0, \tau_0), \tag{8}$$

$$\Lambda \sim IW_p(L_0, t_0), \qquad M \sim Ga(a_1, a_2). \tag{9}$$

The random distribution $H_z(\theta)$ in model (4) is dependent of the level of covariate z. As such, this follows the model proposed by De Iorio et al.

(2004), but our focus is on the application to multivariate data. Matrix R in the model allows for correlation between all components of vector α_i, which implies correlation between different components of the response vector and between different levels of z. The full conditional posterior distributions and details of the posterior simulation scheme are given in the Appendix.

Consider now a general classification approach, and denote the training data set by $y^n = (y_1, \ldots, y_n, x_1, \ldots, x_n, z_1, \ldots, z_n, g_1, \ldots, g_n)$. Here, $y_{iu} = (y_i : g_i = u)$, $u = 1, \ldots, m$, is the response vector for the uth group, $x_{iu} = (x_i : g_i = u)$ is the vector of covariates for fixed effects, $z_{iu} = (z_i : g_i = u)$ is a vector of covariates for random effects and g_i represents the known group label for the ith unit. Consider a new unit for which the response y_{n+1} and covariate vectors x_{n+1} and z_{n+1} are known, but its label g_{n+1} is unknown. We want to assign a label u to the new unit, where $u \in \{1, \ldots, m\}$. Consequently, it is necessary to estimate the classification probability $P(g_{n+1} = u \mid y_{n+1}, y^n)$. Following De la Cruz-Mesía and Quintana (2007) and Gutiérrez et al. (2011), we use

$$P(g_{n+1} = u \mid y_{n+1}, y^n) \approx \frac{1}{C} \sum_{c=1}^{C} \frac{\pi_u p(y_{n+1} \mid \Theta_u^{(c)})}{\sum_l \pi_l p(y_{n+1} \mid \Theta_l^{(c)})}. \tag{10}$$

In (10), $\pi_u = P(g_i = u)$ may be taken as the empirical group proportions. We propose classifying an existing unit, i, and a future one, $n+1$, using the zero–one law considered in Hastie, Tibshirani and Friedman (2001):

$$\begin{aligned} \hat{g}_i &= \arg\max_u P(g_i = u \mid y^n) \quad \text{and} \\ \hat{g}_{n+1} &= \arg\max_u P(g_{n+1} = u \mid y^n, y_{n+1}), \end{aligned} \tag{11}$$

that is, assigning the label as the category that maximizes the classification probability (10). Note that (10) and (11) are part of a generic classification approach which is based on predictive distributions. In particular, (10) relates to the model described in (4) through the $\Theta_u^{(c)}$ quantities, which represent samples from the posterior distribution for model (4), given the training data set y^n. For the wine data analysis later in Section 6, we will let the fixed effects be varieties and random effects be the different regions of origin. For classification purposes, we are assuming that a new wine sample is available, for which we know its anthocyanin concentrations and valley, and the aim is to predict its variety. Note that the variety here corresponds to the label u and there are not other covariates for fixed effects, so the rows on the matrix of a fixed effects are given by $x_i = (1, 0, 0)$ if $g_i = 1$ (Cabernet Sauvignon), $x_i = (0, 1, 0)$ if $g_i = 2$ (Merlot) and $x_i = (0, 0, 1)$ if $g_i = 3$ (Carménère).

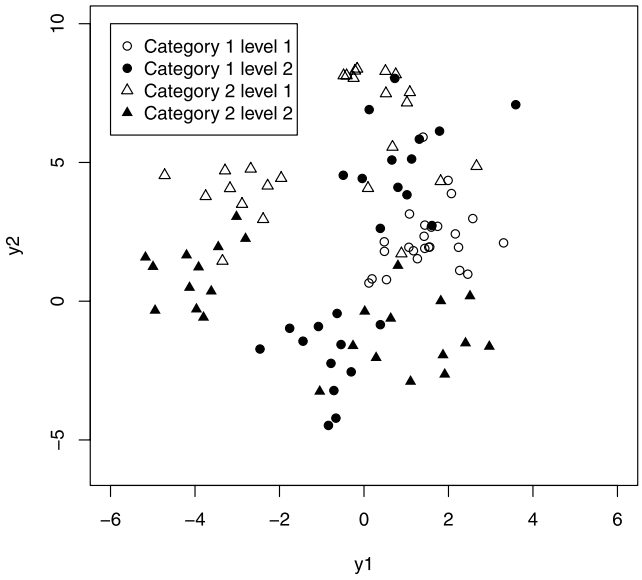

FIG. 1. *Simulated data set.*

5. Classification performance of the proposed model. To evaluate the classification performance of the proposed model, we simulated a data set considering $m = 2$, $n = 100$, $p = 2$, $q = 2$, $k = 2$. We simulated from a mixture of p-variate normal distributions, $\sum_{i=1}^{8} \omega_i N(\mu_i, \Sigma)$, where $\omega_1, \ldots, \omega_8$ are given by (0.25, 0.12, 0.13, 0.1, 0.1, 0.05, 0.12, 0.13), respectively, $\mu_1 = (1.1, 2.3)^t$, $\mu_2 = (0.1, -2)^t$, $\mu_3 = (1.3, 5)^t$, $\mu_4 = (-3, 3.4)^t$, $\mu_5 = (-0.1, 7)^t$, $\mu_6 = (1.8, 5)^t$, $\mu_7 = (-4, 1)^t$, $\mu_8 = (1, -2)^t$ and Σ is given by $\sigma_{11} = 0.932$, $\sigma_{12} = 0.11$ and $\sigma_{22} = 1.632$. Figure 1 shows the simulated data set. Here, $m = 2$ means that we have to classify between two categories with $k = 2$ levels for the covariate z. The hyperparameter values were taken as $\alpha_0 = (0, 0)^t$, $\tau_0 = 100 I_2$, $Q_0 = I_2$, $L_0 = I_2$, $\nu_0 = 4$, $r_0 = 4$, $t_0 = 4$, $R_0 = I_{pk}$, $\gamma_0 = 4$, $\phi_0 = 0.001 I_p$ and $a_1 = a_2 = 1$. Table 1 shows the classification results of the proposed Bayesian semiparametric model (BSP), comparing with linear dis-

TABLE 1
Classification performance for the simulated data set. Values within parenthesis were obtained using leave-one-out cross-validation technique

	BSP		**BP**		**LDA**	
Category	**1**	**2**	**1**	**2**	**1**	**2**
1	47 (43)	3 (7)	47 (35)	3 (15)	42 (42)	8 (8)
2	3 (9)	47 (41)	9 (9)	41 (41)	17 (19)	33 (31)

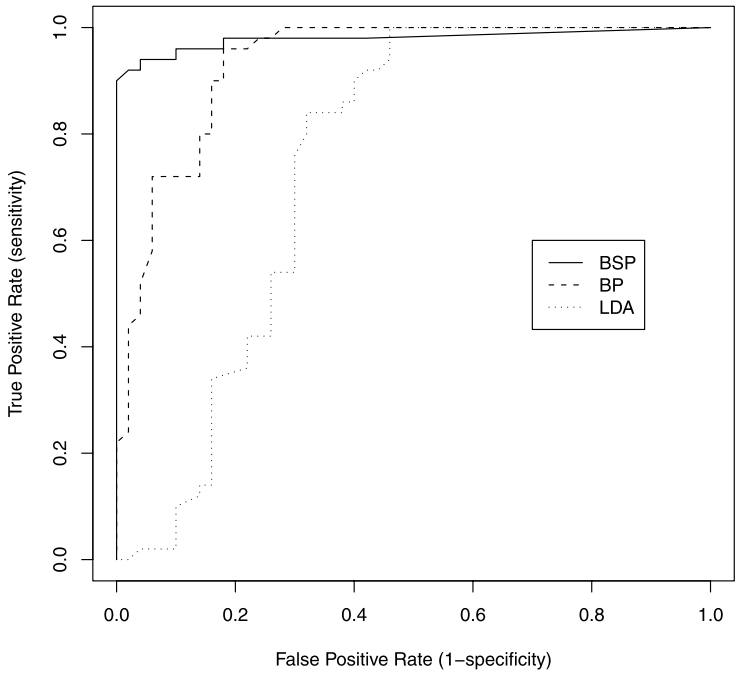

FIG. 2. *ROC curves for classification of the simulated data set under Bayesian semiparametric model (BSP), Bayesian parametric (BP) and linear discriminant analysis (LDA).*

criminant analysis (LDA), which is the usual technique used in the literature for this type of problem, and a parametric (BP) version of model (4), defined as

$$
\begin{aligned}
(y_{iu} \mid x_{iu}, z_{iu}) &\sim N_p(Bx_{iu} + \theta_{iu}, \Sigma_u), \qquad i = 1, \ldots, n, u = 1, \ldots, m, \\
\theta_{iu} &= z_{iu}\alpha + \eta_i, \qquad \eta_i \sim N_p(0, \tau), \\
\alpha &\sim N_{pk}(0, R).
\end{aligned} \tag{12}
$$

Using the proposed BSP model, we obtained a classification error of 7.0% in the training set and 16.0% using leave-one-out cross-validation (LOOCV). In contrast, the BP model resulted in a classification error of 12.0% in the training set and 24.0% under LOOCV, while the corresponding figures for the LDA were 25.0% and 27.0%, respectively. A common way to assess the performance of classification rules is the Receiver Operating Characteristic curve (ROC) shown in Figure 2, which plots the true positive rate against the false positive rate for all the different possible cutpoints. From the ROC curves we also calculated the Area Under ROC curve (AUC) for the three models, with higher values corresponding to models with better discrimination capabilities. We obtained 0.9792 for the BSP model, 0.9334 for the

BP model, and 0.7464 for LDA. These results clearly suggest the superiority of the proposed BSP model for wine authentication in our simulation, compared to the other alternatives.

Another important aspect of the analysis concerns comparing model adequacy of the BP versus our BSP proposal. To this effect, we calculated the Conditional Predictive Ordinates (CPO_i) [Chen, Shao and Ibrahim (2000)], summarized in the log-pseudo marginal likelihood statistic $\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i)$ [Geisser and Eddy (1979)]. Models with higher LPML values are to be preferred. We found the proposed BSP to perform better than the BP, as the corresponding LPML values were -365.3 and -428.7, respectively. As an additional comparison we considered the Deviance Information Criterion (DIC) [Spiegelhalter et al. (2002)]. Models with lower DIC values are to be preferred. It is important to point out that several possible DIC construction are available. Celeux et al. (2006) discuss the DIC for finite mixture and random effects models, giving and comparing different DIC definitions. Our proposal is a mixture model, but the mixture is infinite. Indeed, due to the convolution in (4), the random effects $\{\theta_i\}$ have continuous distributions. Nevertheless, we found some constructions of DIC to yield negative values for the effective dimension p_D, and so we used DIC_1, DIC_2 and DIC_3 of Celeux et al. (2006). The corresponding values were 721.4, 717.3 and 723.1 for the BSP model, and 856.5, 854.4 and 855.4 for the BP model, and the conclusion is the same regardless of the specific construction. In summary, both LPML and DIC criteria consistently point to the superiority of the BSP model compared to the BP one. Overall, the results suggest that the BSP model is more flexible, especially when the data cluster between and within covariate levels.

6. Performance of the model with wine data set. We consider now application of the proposed BSP model to the wine data set. The response vector is formed by the nine anthocyanins listed in Section 2. As covariates, we use grape variety (fixed effects) and valleys (random effects). The hyperparameter values were taken as $\alpha_0 = (0,0,0,0,0,0,0,0,0)^t$, $\tau_0 = 100I_9$, $Q_0 = 0.1I_9$, $L_0 = 0.01I_9$, $\nu_0 = 11$, $r_0 = 65$, $t_0 = 11$, $R_0 = 10I_{pk}$, $\gamma_0 = 11$, $\phi_0 = 0.01I_p$ and $a_1 = a_2 = 1$, where $p = 9$, $q = 3$ and $k = 7$. The resulting prior densities are proper, but the one for B is vague and hence relatively uninformative. The prior density for R is relatively uninformative too. All the prior covariance matrices were assumed of diagonal form. The selected hyperparameter values imply proper but vague prior distributions, representing the lack of genuine prior information on the parameters. To further investigate how sensitive the results are to the above choices, we conducted a prior sensitivity analysis for hyperparameters (a_1, a_2), which control the implied clustering structure, and τ_0, which controls the prior variance for fixed effects. We tried the prior settings listed in the leftmost column of Table 2, which also shows

TABLE 2

Posteriors means and standard deviations (S.D.) for some model quantities. For the combinations indicated in the leftmost column, we present summaries for M, τ and for the fixed effects parameters for Cabernet Sauvignon corresponding to anthocyanins PT (β_{13}) and MV (β_{15})

		PT		MV	
Prior value	**Precision parameter**	**Mean**	**(S.D.)**	**Mean**	**(S.D.)**
$a_1 = a_2 = 0.01$	$M = 1.41$ (0.62)	$\beta_{13} = 2.55$	(0.14)	$\beta_{15} = 4.91$	(0.12)
$\tau_0 = 100 I_9$		$\tau = 0.15$	(0.0136)	$\tau = 0.08$	(0.0088)
$a_1 = a_2 = 1$	$M = 1.65$ (0.57)	$\beta_{13} = 2.68$	(0.09)	$\beta_{15} = 5.07$	(0.11)
$\tau_0 = 100 I_9$		$\tau = 0.16$	(0.0134)	$\tau = 0.08$	(0.0088)
$a_1 = 1$, $a_2 = 0.1$	$M = 1.48$ (0.56)	$\beta_{13} = 2.76$	(0.12)	$\beta_{15} = 5.03$	(0.11)
$\tau_0 = 100 I_9$		$\tau = 0.16$	(0.0129)	$\tau = 0.08$	(0.0081)
$a_1 = 10$, $a_2 = 1$	$M = 3.27$ (0.86)	$\beta_{13} = 2.58$	(0.11)	$\beta_{15} = 4.95$	(0.09)
$\tau_0 = 100 I_9$		$\tau = 0.16$	(0.0126)	$\tau = 0.08$	(0.0079)
$a_1 = a_2 = 1$	$M = 1.43$ (0.50)	$\beta_{13} = 2.46$	(0.17)	$\beta_{15} = 4.93$	(0.15)
$\tau_0 = I_9$		$\tau = 0.16$	(0.0132)	$\tau = 0.08$	(0.0105)
$a_1 = a_2 = 1$	$M = 1.21$ (0.48)	$\beta_{13} = 2.58$	(0.12)	$\beta_{15} = 4.89$	(0.10)
$\tau_0 = 10 I_9$		$\tau = 0.16$	(0.0129)	$\tau = 0.08$	(0.0081)
$a_1 = a_2 = 1$	$M = 1.82$ (0.61)	$\beta_{13} = 2.69$	(0.08)	$\beta_{15} = 5.05$	(0.08)
$\tau_0 = 1000 I_9$		$\tau = 0.17$	(0.0134)	$\tau = 0.09$	(0.0093)
$a_1 = a_2 = 1$	$M = 1.38$ (0.52)	$\beta_{13} = 2.57$	(0.14)	$\beta_{15} = 5.0$	(0.07)
$\tau_0 = 10000 I_9$		$\tau = 0.16$	(0.0125)	$\tau = 0.08$	(0.0082)

the posterior means and standard deviations for some key parameters in the model. Specifically, we show posterior summaries for M, τ and for the fixed effects parameters for Cabernet Sauvignon corresponding to anthocyanins PT and MV. We generally found no notorious changes in these summaries across different prior configurations. Since our emphasis is on classification, we also compared the shape of the predictive distributions and the classification itself for the prior configurations of Table 2. In general, we did not observe big changes in the predictive distributions (data not shown) and the classification was the same for all prior specifications.

Table 3 shows the classification results, where the values within parenthesis were obtained using a LOOCV approach. The classification error obtained in the training set was 0.5%, and 3.2% under LOOCV. These values are better than those obtained by Gutiérrez et al. (2011) with the same data set but applying a Bayesian parametric model, namely, 3.0% in the training set and 3.5% using LOOCV.

Table 4 shows the AUC values, which were calculated based on separate ROC curves for each grape variety, and for each of the BSP and BP mod-

TABLE 3
Misclassification rate for the three grape varieties

Variety	Carménère	C. Sauvignon	Merlot	Error
Carménère	94 (93)	1 (1)	0 (1)	1.1% (2.1%)
C. Sauvignon	0 (0)	228 (225)	0 (3)	0.0% (1.3%)
Merlot	1 (8)	0 (0)	75 (68)	1.3% (10.5%)
Total error				0.5% (3.2%)

TABLE 4
Area under ROC curve

Grape variety	AUC BSM	AUC BPM
Cavernet Sauvignon	0.999999	0.9969221
Merlot	0.999999	0.9967403
Carménère	0.999999	0.9963574

els. All these values are very high, with the BSP model attaining the best performance across the three grape varieties. Comparing the BSP and BP models, we found that the DIC_1, DIC_2, DIC_3, and LPML statistics values were $-5{,}901.5$, $-6{,}368.9$, $-5{,}769.9$ and 2,430.1 for the former, and $-3{,}659.1$, $-5{,}006.6$, $-3{,}446.2$ and 1,351.6 for the latter. Again, these results suggest that the proposed BSP model provides a superior fit, and all the criteria values are consistent. We also note that in all cases, the effective dimension p_D for the DIC was positive.

Figure 3 displays bivariate posterior predictive distributions for Carménère wines from the valleys of Aconcagua, Maipo, Rapel and Curicó, considering the PECU and MVCU anthocyanins. The points on the graph are the observed values. We can see the changes in the posterior predictive distribution across valleys. Predictions for the Aconcagua valley show less variation compared to the Maipo valley. Predictions for The Rapel valley show more variability, with some evidence of asymmetry, as dictated by the observed data, but the model provides a reasonable fit to this behavior. Finally, the Curicó valley also exhibits asymmetry.

Figure 4 shows the bivariate predictive posterior distributions for Cabernet Sauvignon, Carménère and Merlot from the Curicó valley considering the PECU and CY anthocyanins. This plot is interesting because it shows how informative are PECU and CY in terms of the target classification. These two anthocyanins show that some Merlot and Carménère samples have very similar chemical profiles. This behavior is reasonable because some years ago, Carménère, which in other countries disappeared due to phylloxera, was rediscovered in Chile. Formerly, all vineyards planted with this grape variety in Chile were declared as Merlot. Using SSR DNA markers to confirm

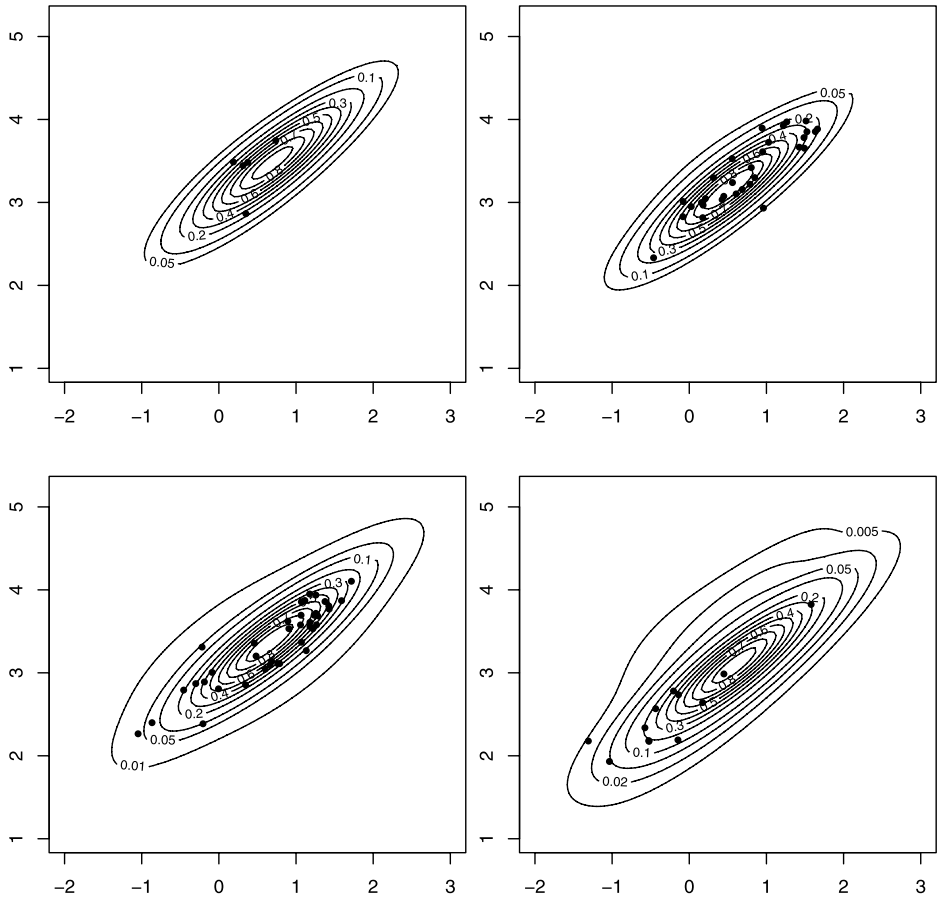

FIG. 3. *Bivariate posterior predictive distributions under the BSP model for Carménère wines from the Aconcagua, Maipo, Rapel and Curicó valleys, with points representing observed values. The anthocyanins considered here were PECU and MVCU in the horizontal and vertical axes, respectively.*

varietal identity, Hinrichsen et al. (2001) found that from a total of 93 vines of five Chilean vineyards, originally planted as Merlot, four vines matched Carménère. This leads to the conclusion that at the time of collecting wine samples, those vineyards declared as Carménère are correctly identified with high probability, but a certain percentage of vineyards declared as Merlot still correspond to Carménère.

7. Concluding remarks. We have proposed a linear mixed effects model for wine authentication, featuring a flexible model for random effects that does not require to restrict ourselves to a given parametric form. We did so by resorting to Dependent Dirichlet Processes, which allow the set of random

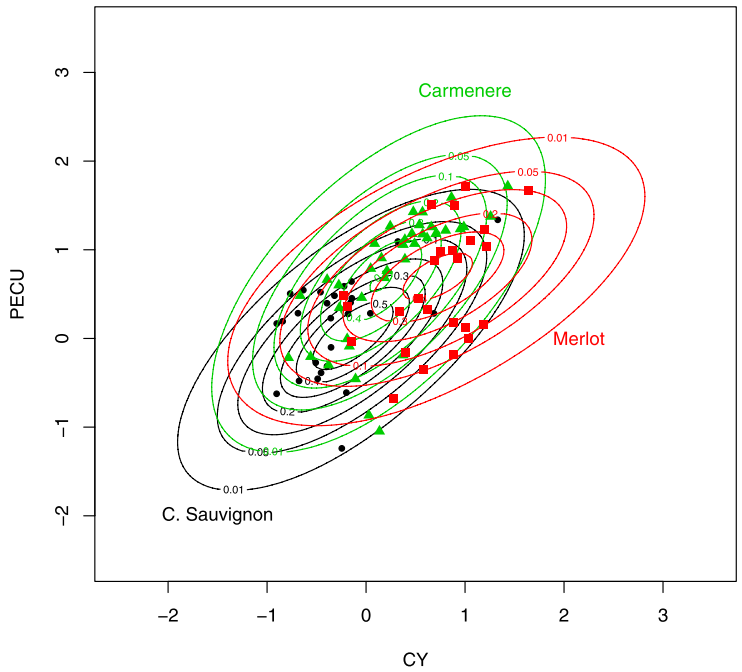

FIG. 4. *Bivariate posterior predictive distributions for Cabernet Sauvignon, Merlot and Carménère wines from the Curicó valley, with points representing the observed values.*

effects distributions to be similar but not identical to each other, depending on values of a set of covariates. For the authentication problem, dependence on covariate levels is important because it is reasonable to think that foods or beverages that come either from the same region of origin, or those which were made with the same technology, could be similar or correlated. The ANOVA-DDP approach was suitable to our purposes, but other types of nonparametric priors could be considered.

The proposed BSP model provided a better fit to the data than a parametric alternative, as we showed in the simulation example and in the application to the wine data. In terms of the target classification, the BSP model also provided slightly better results than other alternatives. Our proposal was motivated by food authentication, but it could be used in any situation where the aim is to classify subjects or units into several groups, on the basis of multiple responses and covariates.

APPENDIX

In this section we give the MCMC algorithm that was used for posterior simulation under the proposed model. Because the model is of conjugate type, we use algorithm 2 in Neal (2000). Let $\mathbf{c} = (c_1, \ldots, c_n)$ denote a vec-

tor that captures the clustering of α_i and let $\alpha = (\alpha_c : c \in \{c_1, \ldots, c_n\})$. To resample the configurations c_i, we proceed with the following two steps:

Step 1. If $c = c_j$ for some $j \neq i$, we compute the probability that the ith element in $\mathbf{c}$ equals other element in the same set as

$$\begin{aligned} &P(c_i = c \mid c_{-i}, \theta_i, \alpha) \\ &\quad = b\frac{n_{-i,c}}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}(\theta_i - z_i\alpha_c)^t\tau^{-1}(\theta_i - z_i\alpha_c)\right\}. \end{aligned} \tag{13}$$

Here $n_{i,c}$ is the number of c_i that are qual to c, c_{-i} are all the c_j for $j \neq i$ and b is such that if $c = c_j$, then $\sum_{j:\, j\neq i}\{P(c_i = c)\} + P(c_i \neq c_j \ \forall j \neq i) = 1$. Next, we compute the probability that c_i is different to any other element in $\mathbf{c}$ as

$$\begin{aligned} &P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, \theta_i, \alpha) \\ &\quad = b\frac{M}{n-1+M}(2\pi)^{-p/2}|\tau|^{-1/2}|R|^{-1/2}|D_i|^{1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}[\theta_i^t\tau^{-1}\theta_i - [\theta_i^t\tau^{-1}z_i]D_i[z_i^t\tau^{-1}\theta_i]]\right\}. \end{aligned} \tag{14}$$

If the imputed value of c_i, sampled based on (13) and (14), is not associated with any other observation, it is necessary to draw a value of α_{c_i} from H_i, the posterior distribution for α based on the prior G_0 and the single observation θ_i. In our case H_i is given by $H_i \equiv N_{pk}(\tilde{\alpha}_i, D_i)$, where $D_i = [z_i^t\tau^{-1}z_i + R^{-1}]^{-1}$, and $\tilde{\alpha}_i = D_i[z_i^t\tau^{-1}\theta_i]$.

Step 2. In the second step, for all $c \in \{c_1, \ldots, c_n\}$ we draw a new value α_c given all the θ_i for which $c_i = c$, that is, from the posterior distribution based on the prior G_0 and all the data points currently associated with latent class c. In our case, this is given by $N_{pk}(\tilde{\alpha}_c, E)$, where $E = [\sum_{i:\, c_i=c} z_i^t\tau^{-1}z_i + R^{-1}]^{-1}$ and $\tilde{\alpha}_c = E[\sum_{i:\, c_i=c} z_i^t\tau^{-1}\theta_i]$.

Now we list all the full conditional distributions for the parametric part of the model. The specific derivation details are straightforward and therefore omitted.

- For fixed effect parameters we have

$$\beta_j \mid \text{other parameters and data} \sim N_p(\tilde{\beta}_j, V_j),$$

where

$$\tilde{\beta}_j = V_j\left[\sum_{u=1}^{g}\left\{\Sigma_u^{-1}\sum_{i=1}^{n_u}\{x_{ij}y_i - x_{ij}x_{il_1}\beta_{l_1} - \cdots - x_{ij}x_{il_q}\beta_{l_q} - x_{ij}\theta_i\}\right\} + \Lambda^{-1}\beta_{0j}\right]$$

and

$$V_j = \left[\sum_{u=1}^{g} \left\{ \sum_{i=1}^{n_u} x_{ij}^2 \Sigma_u^{-1} \right\} + \Lambda^{-1} \right]^{-1},$$

where

$$(l_1, l_2, \ldots, l_q) \neq j, \qquad j = 1, \ldots, q.$$

- For the random effects parameters $\theta_{1u}, \ldots, \theta_{nu}$, $u = 1, \ldots, g$, we have that

$$\theta_{iu} \,|\, \text{other parameters and data} \sim N_p(\tilde{\theta}_{iu}, Q_u), \qquad i = 1, \ldots, n,$$

where

$$Q_u = [\tau^{-1} + \Sigma_u^{-1}]^{-1} \quad \text{and} \quad \tilde{\theta}_{iu} = Q_u[\tau^{-1} z_i \alpha_i + \Sigma_u^{-1} y_i - \Sigma_u^{-1} B x_i].$$

- For hyperparameters $\beta_{01}, \ldots, \beta_{0q}$ we have

$$\beta_{0j} \,|\, \text{other parameters and data} \sim N_p(\tilde{\beta}_{0j}, D_0),$$

where

$$B_{0j} = D_0[\lambda^{-1} \beta_j + \tau_0^{-1} \beta_0], \qquad j = 1, \ldots, q, \quad \text{and} \quad D_0 = [\Lambda^{-1} + \tau_0^{-1}]^{-1}.$$

- For hyperparameter Λ we have

$$\Lambda \,|\, \text{other parameters and data} \sim IW_p(d, E),$$

where

$$E = \sum_{j=1}^{q} (\beta_j - \beta_{0j})(\beta_j - \beta_{0j})^t + L_0 \quad \text{and} \quad d = q + t_0.$$

- Finally, for the covariance matrices $\Sigma_1, \ldots, \Sigma_g$, τ and R we have

$$\Sigma_u \,|\, \text{other parameters and data} \sim IW_p(l_u, H_u),$$

where

$$H_u = \sum_{i=1}^{n_u} (y_i - B x_i - \theta_i)(y_i - B x_i - \theta_i)^t + Q_0 \quad \text{and} \quad l_u = n_u + \nu_0,$$

$$\tau \,|\, \text{other parameters and data} \sim IW_p(s, T),$$

where

$$T = \sum_{i=1}^{n} (\theta_i - z_i \alpha_i)(\theta_i - z_i \alpha_i)^T + \Phi_0 \quad \text{and} \quad s = n + \gamma_0,$$

$$R \,|\, \text{other parameters and data} \sim IW_{pk}(f, O),$$

where

$$O = \sum_{i=1}^{n} \alpha_i \alpha_i^t + R_0 \quad \text{and} \quad f = n + r_0.$$

Acknowledgments. The first author thanks the Comisión Nacional de Investigación Científica y Tecnológica—CONICYT, for supporting his Ph.D. studies at the Pontificia Universidad Católica de Chile.

REFERENCES

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

Berente, B., De la Calle García, D., Reichenbächer, M. and Danzer, K. (2000). Method development for the determination of anthocyanins in red wines by high-performance liquid chromatography and classification of German red wines by means of multivariate statistical methods. *J. Chromatogr. A* **871** 95–103.

Brown, P. J., Fearn, T. and Haque, M. S. (1999). Discrimination with many variables. *J. Amer. Statist. Assoc.* **94** 1320–1329. MR1731493

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *International Conference on Information Fusion*. Florence, Italy.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673 (electronic). MR2282197

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299

De Iorio, M., Johnson, W. O., Müller, P. and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65** 762–771. MR2649849

De la Cruz-Mesía, R. and Quintana, F. (2007). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal. *Biostatistics* **8** 228–238.

De la Cruz-Mesía, R., Quintana, F. A. and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *J. Roy. Statist. Soc. Ser. C* **56** 119–137. MR2359237

Dean, N., Murphy, T. B. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *J. Roy. Statist. Soc. Ser. C* **55** 1–14. MR2224157

Dey, D., Müller, P. and Sinha, D., eds. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics* **133**. Springer, New York. MR1630072

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586

Dunson, D. B., Pillai, N. and Park, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 163–183. MR2325270

Eder, R., Wendelin, S. and Barna, J. (1994). Classification of red wine cultivars by means of anthocyanin analysis. 1st Report: Application of multivariate statistical methods for differentiation of grape samples. *Mitteilungen Klosterneubug* **44** 201–212.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 179–194. MR2268037

Gutiérrez, L., Quintana, F., von Baer, D. and Mardones, C. (2011). Multivariate Bayesian discrimination for varietal authentication of Chilean red wine. *J. Appl. Statist.* **38** 2099–2109.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York. MR1851606

Hinrichsen, P., Narvaez, C., Bowers, J., Boursiquot, J., Valenzuela, J., Muñoz, C. and Meredith, C. (2001). Distinguishing Carmenère from similar cultivars by DNA typing. *American Journal of Enology and Viticulture* **52** 396–399.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G., eds. (2010). *Bayesian Nonparametrics*. Cambridge Univ. Press, Cambridge. MR2722987

Holbach, B., Marx, R. and Ackermann, M. (1997). Bestimmung der anthocyanzusammenset-zung von rotwein mittels hochdruckflüssig chromatographi. *Lebensmittelchemie* **51** 78–80.

Holbach, B., Marx, R. and Ackerman, M. (2001). Bedeutung der shikimisäure und des anthocyanspek-trums für die charakterisierung von rebsorten. *Lebensmittelchenie* **55** 32–34.

Jara, A., Lesaffre, E., Iorio, M. D. and Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Ann. Appl. Statist.* **4** 2126–2149.

MacEachern, S. (1999). Dependent nonparametric processes. In *Proc. Bayesian Statistical Science* 50–55. Amer. Statist. Assoc., Alexandria, VA.

Mafra, I., Isabel, M. P., Ferreira, P., Beatriz, M. and Oliveira, P. (2008). Food authentication by PCR-based methods. *European Food Research Technology* **277** 649–665.

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110. MR2082149

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

OIV (2003). Resolution OENO 22/2003. International Organization of Vine and Wine, Paris.

Otteneder, H., Marx, R. and Zimmer, M. (2004). Analysis of anthocyanin composition of Cabernet Sauvignon and Portugieser wines provides an objective assessment of the grape varieties. *Journal of Grape Wine Research* **10** 3–7.

Otteneder, H., Holbach, B., Marx, R. and Zimmer, M. (2002). Rebsortenbestimmung in Rotwein anhand der Anthocyanspektren. *Mitteilungen Klosterneuburg* **52** 187–194.

Revilla, E., Garcia-Beneytez, E., Cabello, F., Martin-Ortega, G. and Ryan, J. (2001). Value of high-performance liquid chromatographic analysis of anthocyanins in the differentiation of red grape cultivars and red wines made from them. *Journal of Chromatography A* **915** 53–60.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. MR1979380

TOHER, D., DOWNEY, G. and BRENDAN, T. (2007). A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometrics and Intelligent Laboratory Systems* **89** 102–115.

VON BAER, D., MARDONES, C., GUTIÉRREZ, L., HOFMANN, G., BECERRA, J., HITSCHFELD, A. and VERGARA, C. (2005). Varietal authenticity verification of Cabernet sauvignon, Merlot and Carmenère wines produced in Chile by their Anthocyanin, Flavonol and Shikimic acid profiles. *Le Bulletin de L'OIV* **78** 45–57.

VON BAER, D., MARDONES, C., GUTIÉRREZ, L., HOFMANN, G., BECERRA, J., HITSCHFELD, A. and VERGARA, C. (2007). *Anthocyanin, Flavonol, and Shikimic Acid Profiles as a Tool to Verify Varietal Authenticity in Red Wines Produced in Chile. ACS Symposium Series* **952**. American Chemical Society, Washington, DC.

WINTERHALTER, P. (2007). *Authentification of Food and Wine. ACS Symposium Series* **952**. American Chemical Society, Washington, DC.

DIVISIÓN DE BIOESTADÍSTICA
ESCUELA DE SALUD PÚBLICA
FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE
AVENIDA INDEPENDENCIA 939, SANTIAGO
CHILE
E-MAIL: luisgutierrez@med.uchile.cl

DEPARTAMENTO DE ESTADÍSTICA
FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
CASILLA 306, CORREO 22, SANTIAGO
CHILE
E-MAIL: quintana@mat.puc.cl